Novos Caminhos para Aumento de Documentos com templates e Modelos de Linguagem

Lucas Wojcik*, Luiz Coelho†, Roger Granada†, David Menotti*
*Departamento de Informática, Universidade Federal do Paraná, Curitiba, PR, Brasil {lmlwojcik, menotti}@inf.ufpr.br
†unico - idTech, Brasil {luiz.coelho, roger.granada, gustavo.fuhr}@unico.io

***Abstract*—Recent advances in the Natural Language Processing field percolate toward the Document Understanding fields in the manner of new models and tasks, but the topic of data augmentation is often left untouched. This is specially relevant for the document scope, wherein few-shot fine-tuning tasks are of great relevance for most scopes, as properly annotated data is very scarce, and often these systems are used for the very task of annotation. To thrive in these scenarios, we present two new data augmentation techniques where we aim to maximize knowledge using very few instances. One is based on simple structured documents, using templates that symbolize the layout information. The other one uses Large Language Models (LLMs) to rewrite document texts. These methods work only with these two modes, layout and textual. We validate our approaches on the datasets EPHOIE and FUNSD, respectively. Our approach is shown to improve the baseline methods, according to the metrics on simple and joint training.**

***Resumo*—Avanços recentes em processamento de linguagem natural percolam para o campo de reconhecimento de documentos em novos modelos e tarefas, mas o tópico de aumento de dados é raramente discutido. Isto é relevante especialmente para o escopo de documentos, onde tarefas com poucas instâncias de treinamento são de grande importância para muitos domínios, visto que dados bem anotados são escassos, e estes modelos podem ser mesmo utilizados para a própria tarefa de anotação. Para melhorar estes cenários, apresentamos duas novas técnicas de aumento de dados focadas em maximizar o conhecimento de poucas instâncias. Uma é baseada em documentos de estrutura simples, utilizando *templates* que codificam a informação de layout. A outra usa *Large Language Models* (LLMs) para reescreveros textos de documentos. Estes métodos funcionam com dois modos: texto e layout. Validamos nossas técnicas nos *datasets* EPHOIE e FUNSD, respectivamente. Mostramos que nossas técnicas melhoram o *baseline*, de acordo com as métricas para treinamento simples e combinado.**

I. Introdução

Ainda há muito a aprender sobre LLMs (*Large Language Models*) e as possibilidades abertas pelas suas capacidades de generalização com poucos exemplos. A pesquisa em aplicações usando LLMs ainda está num estágio embrionário, especialmente no domínio de reconhecimento de documentos [1]. Algumas técnicas usadas por modelos de processamento de linguagem natural (NLP) foram incorporadas ao domínio de documentos (Exemplos na Seção II), mas os LLMs ainda não são usados.

Ao mesmo tempo, pesquisa em aumento de dados em documentos também parece escassa e específica. Alguns métodos são feitos especificamente para melhorar a performance de alguns modelos, como o SynthDoG do Donut [2], enquanto outros procuram avançar as fronteiras de alguns *datasets* específicos como o DocSynth [3]. Existem até alguns métodos que criam novos *datasets* e problemas como o DocBank [4].

Considerando estas lacunas na literatura, apresentamos algumas técnicas novas para aumento de documentos. Trazemos ao escopo de documentos algumas técnicas desenvolvidas recentemente na pesquisa em NLP. Neste sentido, encontramos melhoras na performance dos modelos de reconhecimento utilizando um LLM para aumento dos textos. Também apresentamos uma técnica simples utilizando *templates* (grafos de entidades) para generalizar o conhecimento de poucas instâncias.

As duas técnicas são bastante versáteis e podem ser usadas para aumentar essencialmente qualquer tipo de documento. Para validá-las, utilizamos os *datasets* FUNSD [5] e EPHOIE [6], com os métodos de LLM e template respectivamente. As duas técnicas são usadas com um modelo bi-modal, usando texto e layout apenas. Até onde sabemos, este é o primeiro trabalho a utilizar LLMs para aumento de *datasets* de documentos, e abre várias possibilidades para a pesquisa futura. Os *datasets* gerados estarão disponíveis publicamente.

O resto deste trabalho consiste das seguintes seções. A Seção II apresenta o panorama geral em reconhecimento de documentos e aumento de dados relacionado. A Seção III apresenta nossa nova metodologia de aumento de documentos, assim como detalha os *datasets* usados e os resultados do nosso aumento. A Seção IV detalha os experimentos feitos para validação das técnicas apresentadas, assim como os resultados obtidos. Finalmente, a Seção V apresenta as conclusões e direções para a pesquisa futura.

II. Trabalhos relacionados

O rápido avanço recente do reconhecimento de documentos se deu graças à incorporação de técnicas de NLP. Isto data desde o LayoutLM [7], cujo *backbone* é baseado no BERT [8]. Inicialmente, BERT foi criado para NLP, mas a estratégia de pré-treinamento utilizada virou uma constante no estado da arte de documentos. Outro exemplo é o GraphDoc [9], cujo mecanismo de atenção se baseia no mecanismo do StarTransformer [10]. A atenção junta as entidades do documento num grafo de acordo com a distância euclidiana entre elas no documento. Em questão de aumento de dados, para documentos há técnicas como DocCutout e DocCutmix [11]. O artigo que as apresenta utiliza as técnicas de Cutout e Mixup já usadas para *datasets* de reconhecimento de imagem para aumentar o *dataset* PubMed [12]. Outro exemplo é o

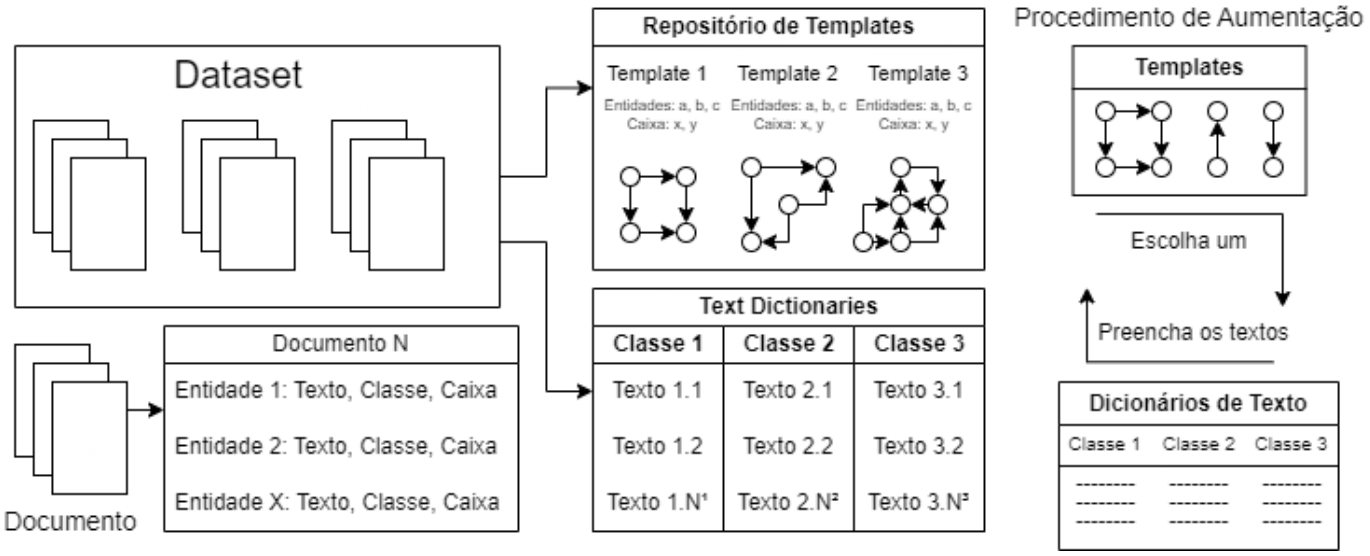

Figura 1. Diagrama representando nossa técnica de aumento de *templates*

SynthDoG [2], que extrai imagens do *dataset* ImageNet [13] para serem o pano de fundo onde texturas de documento são projetadas, com um gerador de texto escrevendo em cima destas texturas. Isto possibilitou aos autores generalizar o modelo de documentos proposto para várias línguas, dado que este não utiliza informação textual.

Algumas GANs para geração de dados também foram usadas para documentos. Um exemplo é DocSynth [3], cujos autores treinaram uma GAN no *dataset* PubLayNet [14], um *dataset* bastante grande de artigos científicos. Uma limitação deste método é a necessidade de grandes quantidades de dados para gerar documentos fidedignos.

DocBank [4] apresenta uma técnica de fraca supervisão para sintetização de documentos. DocBank denomina um *dataset* criado com documentos Word tirados da internet. O modelo utiliza o XML de cada documento para criar anotações com granularidade extremamente fina, a nível de *token*, num *dataset* com meio milhão de instâncias.

Finalmente, [15] apresenta uma extensa lista de técnicas para aumento de documentos. Técnicas para texto são exploradas para aplicação no escopo de documentos legais, como substituição de sinônimos e traduções de ida e volta. LLMs são discutidos (GPT e GPT-2 [16], [17]), mas os autores concluem que, devido à impossibilidade de preservar palavras chave do texto original, eles não podem ser usados. Os autores se limitam a discutir as propostas, sem experimentos.

Este artigo explora dois escopos de documentos e técnicas de aumento descritas na seção seguinte. Com inspiração de alguns artigos de NLP [18], [19], apresentamos novas técnicas para aumento nos *datasets* FUNSD [5] e EPHOIE [6].

III. Metodologia de aumento

Esta Seção detalha nossas técnicas de aumento nos *datasets* selecionados. Usamos o LiLT [20] para os experimentos visto que este modelo é altamente versátil e não utiliza imagens para treino e teste. Isto é importante pois nossas técnicas são *imageless*, não produzem novas imagens.

É possível estender estas técnicas para imagens também, mas nos *datasets* FUNSD e EPHOIE isto é mais difícil pois os documentos são escaneados, e portanto as imagens têm ruído e inclinação e o aumento deveria produzir o mesmo ruído e respeitar a inclinação. EPHOIE [6] é composto por cabeçalhos de provas de escolas chinesas, e FUNSD [5] é composto por formulários em inglês, sendo um subconjunto do RVL-CDIP [21].

Nosso aumento de *templates* é usada para o EPHOIE apenas porque este *dataset* segue o domínio descrito no começo da Seção III-A. Esta técnica pode ser usada para outros *datasets* contanto que o *dataset* seja composto por documentos que sigam tal descrição. Não realizamos aumento de LLM em EPHOIE pois é difícil para os autores validar a saída do modelo em chinês.

Também não realizamos o aumento de *template* no FUNSD pois este *dataset* não segue a descrição de domínio de *template* que elaboramos. Em contraste, o aumento de LLM funciona para este *dataset* pois os textos em inglês são facilmente verificados, e este *dataset* possui vários textos mais complexos que se beneficiam da generalização do LLM. Esta técnica é descrita na Seção III-B.

É importante notar que um documento pode ser definido como uma lista de entidades, onde cada entidade corresponde a um trecho semântico dentro do documento, tal como um título, um parágrafo ou um cabeçalho. Os atributos de cada entidade variam para cada *dataset*, mas neste trabalho lidamos com entidades que possuem, pelo menos, uma *string* (correspondendo ao texto), um conjunto de coordenadas (denotando a posição dentro da imagem) e uma classe.

Nossas técnicas inovam ao olhar para a questão de aumento de documentos como uma questão de texto, algo novo para o campo. Isto faz mais sentido na medida em que os modelos

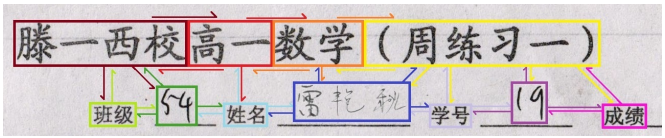

Figura 2. Exemplo de *template* de uma instância do EPHOIE

de estado da arte do campo são adaptações multi-modais de modelos de NLP. A estratégia de *templates* permite uma rápida generalização a partir de uma base de documentos mínima, enquanto a estratégia de LLM permite o aumento de textos bastante complexos. As duas estratégias são complementares e capazes de melhorar o *baseline*, conforme os resultados na Seção IV.

A. Aumento de templates

Alguns tipos de documentos são predefinidos por um conjunto reduzido de *templates*, onde cada documento corresponde a um arranjo específico de entidades dentro da imagem. Definimos um *template* como um grafo direcionado e completo, cujos vértices correspondem às entidades (coordenadas e classe da entidade são atributos do vértice) e cujas arestas possuem um atributo que corresponde a uma de oito direções possíveis (vertical, horizontal e diagonais), indicando a posição relativa entre entidades. A Figura 2 ilustra esta definição. Como indicamos, o grafo é completo, mas esta ilustração suprime várias arestas para manter a legibilidade.

Dados os *templates*, é possível gerar novas instâncias de documentos se for possível gerar textos para cada entidade. Uma técnica simples, que utilizamos para o EPHOIE, é de criar um dicionário de textos do próprio *dataset*, tal que as chaves correspondem às classes de entidades existentes, e cada valor é a lista de todas as *strings* atreladas às entidades com aquela classe dentro do *dataset*. A Figura 1 ilustra esta estratégia. Este método de aumento proposto pode ser resumido nos seguintes passos:

1) Para cada classe, gere uma lista que contenha as strings de todas as entidades do *dataset* com aquela classe.
2) Extraia os *templates* de todos os documentos do *dataset*, unindo *templates* idênticos.
3) Escolha um *template* aleatoriamente e preencha o texto de cada vértice com um texto aleatório da lista de classe.

Além de melhorar a performance do modelo (de acordo com os resultados na Seção IV), esta estratégia também possibilita a geração de uma grande variedade de instâncias usando uma quantidade mínima de exemplos, dado que os *templates* sejam conhecidos. Outro domínio para aplicação é o de identidades. Com dicionários de nomes, órgãos expedidores e cidades (facilmente encontrados na internet), uma única instância (que não precisa ser anotada) de cada órgão possibilita a criação de um *dataset* bastante rico.

B. Aumento de LLM

Para o FUNSD, que possui textos e *templates* muito mais complexos, experimentamos com algumas técnicas inspiradas em avanços recentes na pesquisa em NLP para aumento de dados [18], [19]. Utilizamos a excelente capacidade dos LLMs de geração de dados com poucos exemplos para sintetizar novas opções para cada documento. Nossa técnica está ilustrada na Figura 3.

Como o FUNSD possui também uma grande variedade de tipos de texto, não foi possível encontrar uma técnica que contemplasse todos os casos do *dataset*. Classificamos todas as entidades em quatro classes de acordo com o tipo de texto. Estas classes são: sentenças complexas, questões simples (entidades com textos como ”name”, ”R&D”), respostas simples (nomes, datas) e nenhum (como strings com zero a dois caracteres ou códigos complexos).[1] Utilizamos uma técnica diferente para cada tipo de entidade.

Para as sentenças complexas, pedimos para o LLM reescrever o texto até cinco vezes. O LLM substitui vários termos e altera a sintaxe, aumentando bastante a variedade do *dataset*. Para as perguntas simples, pedimos para o LLM gerar uma lista de sinônimos. O número de gerações varia por entidade nos dois casos. Apesar de pedirmos um número específico de gerações no primeiro caso, nem sempre o LLM gera sentenças relevantes. Por este motivo e devido ao fato do modelo adicionar um preenchimento indesejado[2], o resultado passa por uma etapa de curadoria humana para extrair apenas os resultados relevantes.

As respostas simples são aumentadas através de algumas técnicas mais simples. Nomes são substituídos usando um dicionário de nomes estadunidense, iniciais são expandidas para nomes completos e nomes completos são retraídos em iniciais. Também mudamos formatos como ”Sobrenome, Nome”para ”Nome, Sobrenome”e vice-versa. Datas são geradas mudando o formato: ”MM-DD-YYYY”pode virar ”Month DD, YYYY”e vice-versa. Números e medidas são aumentados com ruído de OCR, com alguns dígitos sendo substituídos de forma aleatória. O quarto tipo de texto não é aumentado.

A técnica de LLMs pode ser utilizada para a maioria dos *datasets*, em especial aqueles cujos documentos são compostos por textos mais complexos como o FUNSD. Esta é a primeira tentativa na literatura de utilizar um LLM para aumento de dados em tarefas de reconhecimento de documentos, até onde sabemos. Além de melhorar a performance do modelo, esta técnica também abre novos caminhos para a pesquisa futura. Com alguns avanços recentes em gerações de imagem [22], esta técnica poderia ser usada para aumentar imagens também, possibilitando que modelos tri-modais (texto, layout e imagens) [23] sejam utilizados.

C. Descrição dos conjuntos de dados

A Tabela I apresenta as quantidades de instâncias e entidades para os *datasets* EPHOIE e FUNSD. Para EPHOIE, aumentamos o número de documentos usando a técnica de *templates*. De 1183 documentos no treino, extraímos 1046

[1]Estas classes não necessariamente correspondem às quatro classes predefinidas pelo próprio *dataset*.

[2]Textos de *chatbot* como: “Certainly! Here’s a list of synonyms:”, ”Thank you for asking! Here’s some synonyms:”, etc.

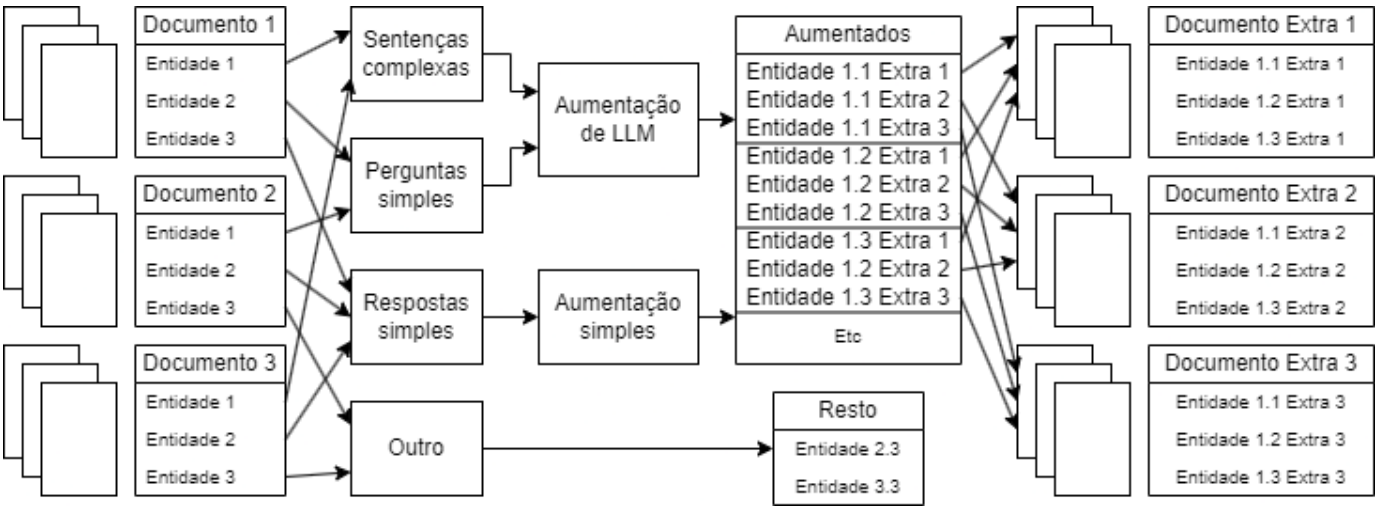

Figura 3. Diagrama representando o aumento de LLM

Tabela I
NÚMERO DE INSTÂNCIAS E ENTIDADES EM CADA PARTIÇÃO

	FUNSD		EPHOIE	
Partição	Documentos	Entidades	Documentos	Entidades
Treino (real)	149	7411	1183	12411
Teste	50	2332	311	3343
1 Sintético	149	7411	1200	12921
2 Sintéticos	298	14822	2400	25850
3 Sintéticos	447	22233	3588	38656
4 Sintéticos	696	29644		
5 Sintéticos	745	37055		

templates únicos, dos quais 249 são subgrafos de outros *templates*. A Tabela II mostra o número de entidades no conjunto de treino do EPHOIE. Como visto, o número de entidades é desbalanceado.

Como uma regra, as partições subsequentes contém as anteriores. Isto significa que a partição de dois aumentos do FUNSD contém todos os 149 documentos da partição de uma, mais 149 gerações e assim por diante. Para o FUNSD, nosso aumento produz uma "cópia" de cada documento na partição real para cada aumento. Portanto, a partição de treino de um aumento possui 298 documentos: os 149 do *dataset* original mais uma geração sintética para cada documento (com os textos substituídos). Isto quer dizer que o número de entidades permanece o mesmo entre documento real e aumento.

Para o FUNSD, usamos o Llama-2-7b-hf [24] pré-treinado para os aumentos de texto. As estatísticas de entidade são apresentadas na Tabela III. Também apresentamos o número de aumentos por entidade e tipo de aumento. Conforme mostrado, FUNSD também possui um desequilíbrio entre entidades, tanto nas classes originais quanto nossa classificação por tipo de texto. As entidades sem aumentos correspondem a textos inválidos como strings vazias ou com até dois caracteres (caixas de seleção, falhas de OCR) ou textos que o LLM falhou em compreender, como longos códigos de letras e número, acrônimos e compostos químicos.

Tabela II
NÚMERO DE INSTÂNCIAS POR CLASSE NO TREINO REAL DO EPHOIE

Entity type	Amount
Other	5679
Exam Number	128
Score	377
Name	2365
Student Number	422
School	1358
Grade	441
Seat Number	184
Class	1625
Subject	376
Candidate Number	467
Test Time	79

Tabela III
NÚMERO DE ENTIDADES NO TREINO REAL DO FUNSD

Número de Entidades		Número de Gerações		Por Tipo	
Header	411	1	978	Complexo	647
Question	3266	2	1580	Sinônimo	4106
Answer	2802	3	1560	Simples	375
Other	902	4 ou mais	1010	Nenhum	2283
Total	7411	Gerações	14611	Aumentados	5128

IV. EXPERIMENTAÇÃO

A. Modelo e protocolos

Para validar nossas técnicas, usamos LiLT [20], um transformer [25] bi-modal. Este modelo é composto por dois fluxos de transformer independentes que se conectam através de um mecanismo de atenção chamado "complementação de atenção bi-direcional" que substitui o modelo clássico de atenção do transformer original. Cada fluxo corresponde a uma modalidade: texto e layout, sendo que o fluxo de texto pode corresponder a um modelo de NLP qualquer. Esta arquitetura está detalhada no artigo do LiLT.

Este design permite ao LiLT funcionar com vários modelos de texto de transformer diferentes, incluindo modelos treinados em línguas diferentes. Isso permite aos autores testarem no

XFUND [26], um *dataset* multilíngue que pode ser considerado uma extensão do FUNSD. LiLT utiliza tanto um RoBERTa [27] treinado em inglês para o FUNSD quando o InfoXLM [28], um modelo multilíngue, para FUNSD e XFUND. Os autores também utilizam o InfoXLM e um RoBERTa chinês [29] para o EPHOIE.

Optamos por não treinar o modelo no XFUND pois é difícil para os autores validar as gerações do LLM em algumas das línguas deste *dataset*, como chinês e japonês. Focamos no FUNSD e EPHOIE, utilizando os modelos LiLT-RoBERTa-EN e LiLT-InfoXLM, que foram disponibilizados pelos autores. [3]

B. Experimentos e resultados

Nos nossos experimentos, adicionamos nossas gerações sintéticas ao conjunto de treinamento do *dataset* original. Nosso *baseline* é a performance reportada pelo artigo do LiLT. Realizamos o fine-tuning de cada modelo na tarefa de reconhecimento de entidades (*Semantic Entity Recognition* - SER), que é a tarefa de assinalar a classe correta para cada entidade no documento. Para o FUNSD, também realizamos a tarefa de extração de relações (*Relation Extraction* - RE), que é a tarefa de estabelecer as relações entre *headers* e *questions*, e *questions* e *answers* de acordo com as anotações do documento.

Para cada tarefa, algumas camadas extras de adaptação foram adicionadas ao modelo base para realizar a classificação. Uma explicação completa do design destas camadas está no artigo do LiLT. Elas estão implementadas no GitHub oficial dos autores do LiLT, e utilizamos esta implementação nos nossos experimentos.

Os resultados para o FUNSD estão apresentados na Tabela IV. Os autores do LiLT apresentam um resultado melhor para a tarefa de SER com o modelo RoBERTa-EN (o resultado para o InfoXLM é de 85.86), mas para a tarefa de RE não há um resultado usando o RoBERTa. Isto pode ser porque esta tarefa é utilizada para comparação com o *dataset* multilíngue XFUND. Para melhor comparar nossos resultados, treinamos SER com o RoBERTa-EN e RE com o InfoXLM. Ambas as tarefas são realizadas no cenário monolíngue (ou seja, não utilizamos as outras partições do XFUND no treinamento já que trabalhamos apenas com o FUNSD.)

Conforme a Tabela IV, nossa técnica melhora o *baseline*, melhorando a faixa de erro em 1.35 para SER. Também melhoramos na tarefa de RE por 7.76. A razão para a maior melhora em RE pode ser o fato de que o conhecimento do InfoXLM da língua inglesa é expandido com o nosso aumento. Isto é relevante pois InfoXLM é um modelo mais geral, pré-treinado num cenário multilíngue. Além disso, nossos aumentos mostram uma melhora consistente em todos os cenários de treinamento.

Na Tabela V, mostramos os resultados para a tarefa SER no *dataset* EPHOIE. Como *baseline*, utilizamos os resultados reportados pelo LiLT. Os autores treinam em EPHOIE utilizando a mesma estratégia do FUNSD com a camada de adaptação. Já que não temos acesso ao LiLT-RoBERTa-ZH usado no resultado reportado, utilizamos o LiLT-InfoXLM. Mesmo com um modelo mais geral, conseguimos melhorar a performance por mais de metade da margem de melhora possível. Levamos o *baseline* de 97.97 para 99.2, uma melhora de 1.23 de possíveis 2.03.

[3]O modelo com o RoBERTa chinês não está no GitHub official do LiLT. No entanto, foi possível melhorar o resultado do EPHOIE utilizando o InfoXLM, um modelo mais geral.

Tabela IV
RESULTADOS PARA SER E RE NO FUNSD. REPORTAMOS O *micro-averaged* F1 NO TESTE

Partição	SER (RoBERTa-EN)	RE (InfoXLM)
Apenas real (Reportado)	88.41	62.76
Real + 1 sintético	88.82	64.4
Real + 2 sintéticos	**89.76**	68.44
Real + 3 sintéticos	89.04	**70.52**
Real + 4 sintéticos	<u>89.72</u>	<u>70.35</u>
Real + 5 sintéticos	89.02	69.55

Tabela V
RESULTADOS PARA SER EM EPHOIE. REPORTAMOS O *micro-averaged* F1 NO TESTE. NOSSOS RESULTADOS USAM O MODELO INFOXLM

Partição	F1 do teste
Apenas real (Reportado) - RoBERTa-ZH	97.97
Apenas real (Reportado) - InfoXLM	97.59
Real + 1 sintético	**99.2**
Real + 2 sintéticos	<u>99.19</u>
Real + 3 sintéticos	99.13

V. CONCLUSÃO E TRABALHO FUTURO

Neste trabalho, apresentamos duas novas estratégias de aumento, com foco no escopo de escassez de documentos e/ou anotações. Estas técnicas exploram novos caminhos para utilizar a informação contida nos próprios documentos e no grande amálgama de conhecimento dos LLMs. Nossas técnicas mostram uma melhora consistente na performance do modelo treinado em dois *datasets*: FUNSD e EPHOIE. Utilizando inspiração provinda de pesquisa em campos adjacentes, apresentamos novas maneiras de pensar sobre aumento de documentos.

Estes resultados podem ser aprofundados com pesquisa em aplicabilidade para outros escopos de documentos. Existem vários *datasets* com problemas em aberto que se beneficiariam das técnicas apresentadas. Apesar de uma e outra técnica serem limitadas no quesito aplicabilidade, a interseção entre essas limitações (conforme explicado na Seção III) não é muito grande, de forma que há espaço para tratar de vários escopos a mais além dos *datasets* apresentados neste trabalho.

Em termos puramente técnicos, nossos métodos são limitados, no caso do LLM, pela necessidade de curadoria humana, e no caso dos *templates*, pela necessidade de geração de texto. Uma linha de trabalho futuro é o *fine-tuning* dos LLMs para melhor formatação da saída, bem como para geração de dados a partir dos nomes das classes e informações gerais sobre o tipo de texto. Isto elimina as limitações observadas nos dois métodos.

AGRADECIMENTOS

Os autores gostariam de agradecer UNICO por todo o apoio durante a elaboração deste projeto de pesquisa, e também NVIDIA Corporation pela generosa doação da GPU Quadro RTX 8000 que tornou nossos experimentos possíveis. David Menotti agradece ao CNPq (# 315409/2023-1).

REFERÊNCIAS

[1] J. Kaddour, J. Harris, M. Mozes, H. Bradley, R. Raileanu, and R. McHardy, "Challenges and applications of large language models," *ArXiv*, vol. abs/2307.10169, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:259982665

[2] G. Kim, T. Hong, M. Yim, J. Nam, J. Park, J. Yim, W. Hwang, S. Yun, D. Han, and S. Park, "Ocr-free document understanding transformer," in *European Conf. on Computer Vision (ECCV)*, 2022.

[3] S. Biswas, P. Riba, J. Lladós, and U. Pal, "Docsynth: A layout guided approach for controllable document image synthesis," in *Int. Conf. on Document Analysis and Recognition (ICDAR)*, 2021.

[4] M. Li, Y. Xu, L. Cui, S. Huang, F. Wei, Z. Li, and M. Zhou, "Docbank: A benchmark dataset for document layout analysis," 2020.

[5] J.-P. T. Guillaume Jaume, Hazim Kemal Ekenel, "Funsd: A dataset for form understanding in noisy scanned documents," in *Accepted to ICDAR-OST*, 2019.

[6] J. Wang, C. Liu, L. Jin, G. Tang, J. Zhang, S. Zhang, Q. Wang, Y. Wu, and M. Cai, "Towards robust visual information extraction in real world: New dataset and novel solution," in *Proceedings of the AAAI Conf. on Artificial Intelligence*, 2021.

[7] Y. Xu, M. Li, L. Cui, S. Huang, F. Wei, and M. Zhou, "Layoutlm: Pre-training of text and layout for document image understanding," *CoRR/arXiv*, vol. abs/1912.13318, 2019.

[8] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Conf. of the North American Chapter of the Association for Computational Linguistics*, Jun. 2019, pp. 4171–4186.

[9] Z. Zhang, J. Ma, J. Du, L. Wang, and J. Zhang, "Multimodal pre-training based on graph attention network for document understanding," *CoRR/arXiv*, vol. abs/2203.13530, 2022.

[10] Q. Guo, X. Qiu, P. Liu, Y. Shao, X. Xue, and Z. Zhang, "Star-transformer," in *Conf. of the North American Chapter of the Association for Computational Linguistics*, jun 2019.

[11] Y. Lee, T. Hong, and S. Kim, "Data augmentations for document images," in *SDU@AAAI*, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:232369460

[12] K. Li, C. Wigington, C. Tensmeyer, H. Zhao, N. Barmpalios, V. I. Morariu, V. Manjunatha, T. Sun, and Y. Fu, "Cross-domain document object detection: Benchmark suite and method," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 12 915–12 924.

[13] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Int. Conf. on Neural Information Processing Systems (NeurIPS)*, 2012, pp. 1097–1105.

[14] X. Zhong, J. Tang, and A. J. Yepes, "Publaynet: largest dataset ever for document layout analysis," in *2019 Int. Conf. on Document Analysis and Recognition (ICDAR)*. IEEE, Sep. 2019, pp. 1015–1022.

[15] C. Márk and T. Orosz, "Comparison of data augmentation methods for legal document classification," *Acta Technica Jaurinensis*, vol. 15, 07 2021.

[16] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding by generative pre-training," 2018.

[17] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," 2019. [Online]. Available: https://www.semanticscholar.org/paper/Language-Models-are-Unsupervised-Multitask-Learners-Radford-Wu/9405cc0d6169988371b2755e573cc28650d14dfe

[18] J. Ye, N. Xu, Y. Wang, J. Zhou, Q. Zhang, T. Gui, and X. Huang, "Llm-da: Data augmentation via large language models for few-shot named entity recognition," 2024.

[19] Z. Guo, P. Wang, Y. Wang, and S. Yu, "Improving small language models on pubmedqa via generative data augmentation," 2023.

[20] J. Wang, L. Jin, and K. Ding, "LiLT: A simple yet effective language-independent layout transformer for structured document understanding," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, May 2022, pp. 7747–7757. [Online]. Available: https://aclanthology.org/2022.acl-long.534

[21] A. W. Harley, A. Ufkes, and K. G. Derpanis, "Evaluation of deep convolutional nets for document image classification and retrieval," in *International Conference on Document Analysis and Recognition (ICDAR)*, 2015.

[22] P. Dhariwal and A. Nichol, "Diffusion models beat gans on image synthesis," in *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, Eds., vol. 34. Curran Associates, Inc., 2021, pp. 8780–8794. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2021/file/49ad23d1ec9fa4bd8d77d02681df5cfa-Paper.pdf

[23] Y. Huang, T. Lv, L. Cui, Y. Lu, and F. Wei, "Layoutlmv3: Pre-training for document ai with unified text and image masking," *CoRR/arXiv*, vol. abs/2204.08387, 2022.

[24] H. Touvron *et al.*, "Llama 2: Open foundation and fine-tuned chat models," *ArXiv*, vol. abs/2307.09288, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:259950998

[25] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, ser. NIPS'17. Red Hook, NY, USA: Curran Associates Inc., 2017, p. 6000–6010.

[26] Y. Xu, T. Lv, L. Cui, G. Wang, Y. Lu, D. Florencio, C. Zhang, and F. Wei, "XFUND: A benchmark dataset for multilingual visually rich form understanding," in *Findings of the Association for Computational Linguistics: ACL 2022*. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 3214–3224. [Online]. Available: https://aclanthology.org/2022.findings-acl.253

[27] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," 2019, cite arxiv:1907.11692. [Online]. Available: http://arxiv.org/abs/1907.11692

[28] Z. Chi, L. Dong, F. Wei, N. Yang, S. Singhal, W. Wang, X. Song, X.-L. Mao, H. Huang, and M. Zhou, "InfoXLM: An information-theoretic framework for cross-lingual language model pre-training," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Online: Association for Computational Linguistics, Jun. 2021, pp. 3576–3588. [Online]. Available: https://aclanthology.org/2021.naacl-main.280

[29] Y. Cui, W. Che, T. Liu, B. Qin, S. Wang, and G. Hu, "Revisiting pre-trained models for Chinese natural language processing," in *Findings of the Association for Computational Linguistics: EMNLP 2020*, T. Cohn, Y. He, and Y. Liu, Eds. Online: Association for Computational Linguistics, Nov. 2020, pp. 657–668. [Online]. Available: https://aclanthology.org/2020.findings-emnlp.58